Research Article | Vol. 15, No. 1 / 1 Jan 2024 / Biomedical Optics Express 460
Biomedical Optics EXPRESS

Pose estimation via structure-depth information from monocular endoscopy images sequence

SHIYUAN LIU,[1,2] JINGFAN FAN,[1,6,†] LIUGENG ZANG,[1] YUN YANG,[3] TIANYU FU,[4] HONG SONG,[5] YONGTIAN WANG,[1] AND JIAN YANG[1,7,†]

[1]*Beijing Engineering Research Center of Mixed Reality and Advanced Display, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China*
[2]*China Center for Information Industry Development, Beijing 100081, China*
[3]*Department of General Surgery, Beijing Friendship Hospital, Capital Medical University; National Clinical Research Center for Digestive Diseases, Beijing 100050, China*
[4]*Institute of Engineering Medicine, Beijing Institute of Technology, Beijing 100081, China*
[5]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China*
[6]*fjf@bit.edu.cn*
[7]*jyang@bit.edu.cn*
[†]equal

Abstract: Image-based endoscopy pose estimation has been shown to significantly improve the visualization and accuracy of minimally invasive surgery (MIS). This paper proposes a method for pose estimation based on structure-depth information from a monocular endoscopy image sequence. Firstly, the initial frame location is constrained using the image structure difference (ISD) network. Secondly, endoscopy image depth information is used to estimate the pose of sequence frames. Finally, adaptive boundary constraints are used to optimize continuous frame endoscopy pose estimation, resulting in more accurate intraoperative endoscopy pose estimation. Evaluations were conducted on publicly available datasets, with the pose estimation error in bronchoscopy and colonoscopy datasets reaching 1.43 mm and 3.64 mm, respectively. These results meet the real-time requirements of various scenarios, demonstrating the capability of this method to generate reliable pose estimation results for endoscopy images and its meaningful applications in clinical practice. This method enables accurate localization of endoscopy images during surgery, assisting physicians in performing safer and more effective procedures.

1. Introduction

Pose estimation based on endoscopic images has been widely applied in clinical surgeries [1]. Surgeons evaluate the spatial relationship of the surgical environment and measure the distance between surgical instruments and the surgical surface based on their experience [2]. However, the tunnel view in endoscopic images may require multiple observations by the surgeon to obtain information about the same scene, which increases the time and risk of the surgery [3]. Surgeons' spatial perception of the surgical direction and distance relies on the feedback mechanism generated by the actual endoscope motion and image [4]. These feedback mechanisms help surgeons avoid accidental contact with critical tissues and organs during the surgery [5]. Therefore, utilizing the image structure and depth information in consecutive frames of endoscopic images to estimate endoscope motion and pose plays a crucial role in enabling surgeons to identify 3D structures during the surgery.

There are three primary strategies for endoscopy pose estimation, namely magnetic pose estimation based on an electromagnetic system, optical pose estimation based on stereoscopic vision, and image pose estimation based on motion prediction and registration [6,7,8]. Magnetic pose estimation involves integrating magnetic sensors into the endoscope to record its position as it passes through an artificial magnetic field during surgery [9]. This method has been applied in

#498262 https://doi.org/10.1364/BOE.498262
Journal © 2024 Received 16 Jun 2023; revised 8 Dec 2023; accepted 14 Dec 2023; published 22 Dec 2023

Biomedical Optics EXPRESS

puncture surgery and bronchoscopy [10,11]. However, most electromagnetic localization systems have weak anti-interference capabilities and cannot be used in conjunction with MRI devices. The presence of ferromagnetic surgical instruments used during surgery can distort the magnetic field, leading to inaccurate localization [12,13]. Additionally, the electromagnetic localization system is bulky, making it challenging to integrate the electromagnetic sensors onto certain endoscopes [14]. Optical pose estimation involves attaching infrared reflective markers on the endoscope and tracking its movement by locating the markers. This method has been applied in the navigation system for rhinoscopy [15,16]. However, this system requires the marker's position to be fixed relative to the endoscope, and occlusion of the marker can lead to positioning failure, which can limit the surgeon's ability to operate [17,18].

The image position method relies solely on intraoperative endoscopy images and does not require any additional hardware equipment [19]. Research on this method can be broken down into three primary tasks: endoscopy location initialization, interframe image motion estimation, and endoscopy pose optimization [20,21,22]. To initialize the endoscopy location, a series of virtual images from CT data are taken, and the location system is initialized by identifying the coordinates of the most similar virtual images through a similarity calculation between the real endoscopy image and the virtual image [23,24,25]. Other visual features, such as edge features, have also been used to compute image similarity. However, this approach may result in the loss of frames in continuous frame pose estimation due to the presence of multiple similar image features [26,27,28]. Networks have been employed to produce real texture and depth maps of the endoscope, and similarity features between images are compared using depth information. However, this method cannot achieve real-time positioning. Banach et al. [29] developed a deep learning model for feature extraction based on generative adversarial networks, which can improve localization performance in weak-textured scenes by leveraging the diversity of perceptual information. However, this method cannot achieve real-time localization. Mahmoud et al. [30] utilized the Structure from Motion (SfM) algorithm to reconstruct 3D structures from videos and used their geometric features for localization. However, similar structures in surgical scenes, such as intestines and bronchi, may cause positioning errors. Zhao et al. [31] achieved endoscopy positioning through the mutual constraints of frame motion and structural prior information. Nevertheless, during experiments, there was a risk that the initial position might not be within the motion range.

Regarding image interframe motion estimation, the motion estimation algorithm for endoscopy images is similar to Simultaneous Localization and Mapping (SLAM) [32]. Some researchers have utilized sparse feature matching to estimate precise clustering frame locations. They segmented video frames based on the parallax criterion and used the variational method, combining the zero-mean normalized cross-correlation and gradient norm regularization, to estimate the endoscopy position [33,34,35]. In feature geometry constraint, an unsupervised deep learning motion method has been employed to estimate the 3D point projection constraint relation between front and back frames. However, the method's applicability is affected by the extraction of corresponding points during training [36,37]. In a scenario of constant illumination, visual features have been used to estimate the direction of motion, and an optical flow method has been incorporated to track endoscopy motion, leading to good results [38,39]. In capsule endoscopy with a wide range of motion, Dimas et al. [14] employed visual features and depth information, combined with an optical flow tracking method, for surgical assistance. This enables the realization of the visualization and localization functions of wireless capsule endoscopy in the intestinal tract. However, relying on image visual features for localization results can be influenced by the image quality, particularly in weak-textured images.

In regards to endoscopy pose optimization, the gradient and photometric-based methods might be affected by image quality, while the depth information-based methods possess stronger robustness in the optimization process by reflecting the structural information of the current

Biomedical Optics EXPRESS

position [40,41]. Recasens et al. [42] utilized self-supervised deep networks to generate pseudocolor-depth images and optimize the endoscopy position using photometric residuals. However, in intraoperative data with random motion direction, estimation results may exhibit position drift, leading to the planarization of the 3D reconstruction structure and the loss of some spatial information [43,44]. Lei et al. [45] employed the Shape from Shading (SfS) algorithm to generate the corresponding depth map of the image and obtain the target shape for localization optimization from the single-intensity image. Visual-based methods can track and locate textured images [46], while learning-based methods can perceive more diverse information, even with texture-less images [47].

This paper proposes a novel method for pose estimation of a monocular endoscopic image sequence utilizing structure-depth information. By utilizing the structural differences and depth information of endoscopic images, the joint estimation of endoscope motion posture is achieved, and the adaptive cavity boundary constraint method is employed to suppress the cumulative error of motion between sequence frames. The contributions of this study can be summarized into three main aspects. Firstly, the structure difference vector generation (ISD) network is constructed by using the structure difference information between images, and the range of the initial frame is limited by the region difference information of successive frame segments. Secondly, the depth information entropy similarity measurement strategy was introduced to convert the endoscopic pose positioning into the structure-depth information matching between images. Thirdly, the attitude estimation of continuous frame endoscope is optimized. The geometric features are used to extend the three-dimensional space domain and reduce the influence of frame sequence cumulative error on the attitude estimation accuracy.

2. Methods

Accurately estimating the pose of endoscopy is crucial for constructing an internal 3D surface model that provides feedback on intraoperative endoscopy movement, ensuring a successful operation for the surgeon. In this section, we detail the method of training ISD networks to produce image structure difference vectors for initial frame location. We elaborate on how to achieve precise position information from depth estimation outcomes and introduce a relocation method for optimizing pose estimation results. The overall training architecture is depicted in Fig. 1, where all of the concepts we describe in this section are introduced.

2.1. Flowchart of the method

The method's flowchart is depicted in Fig. 1. In the initial frame location module, the cavity region reconstructed from CT data and corresponding endoscopy image sequences are divided into several subsections. Then, the ISD network training generates the structure difference vector of endoscopy images. Subsequently, the similarity measure of the structure difference vector between the initial frame of intraoperative endoscopy and the sequence frame of virtual endoscopy is used to locate the initial frame of the intraoperative endoscopy. In the pose estimation module, during the filming process of virtual endoscopy, the pose and depth information of each frame are obtained. Then, the pose estimation of intraoperative endoscopy images is achieved by maximizing the similarity measure of depth information entropy between intraoperative and virtual endoscopy images. In the repositioning optimization module, the cavity boundary constraint is constructed with the position of the current frame. If the current frame's position exceeds the adaptive error threshold, the current frame is repositioned as the new initial frame.

2.2. Initial frame location

To obtain the initial frame location information, the ISD network is constructed as depicted in Fig. 2. In the input preparation step, the cavity's surface is segmented using preoperative CT data, and the cavity's internal structure is imaged using virtual endoscopy. Then, the intraoperative and

Biomedical Optics EXPRESS

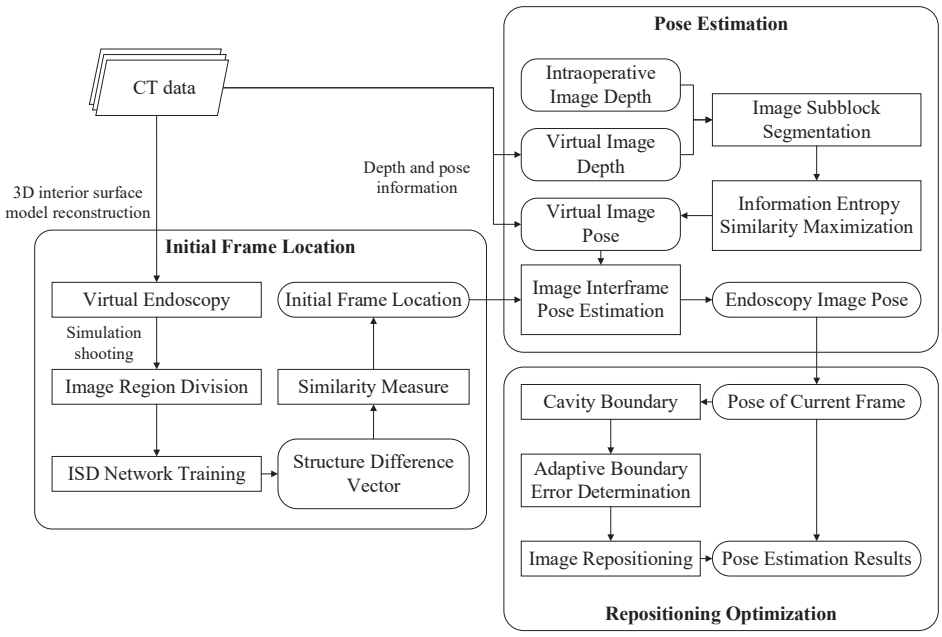

Fig. 1. Flowchart of the proposed structure-depth information based method for pose estimation from monocular endoscopy image sequence method.

virtual endoscopy image sequences are segmented into different regional structures by surface region structure division. A virtual image and an intraoperative image in the same segment are chosen as positive sample and reference images, respectively, while a virtual image in a different segment is utilized as a negative sample image. The three images are simultaneously inputted into the ISD network. Based on the triple network structure, the network carries out image structure difference vector training using three similar branch network structures.

In the branch network (indicated by the dotted box in Fig. 2), the input of the convolutional coding layer is a 3×H × W image, where 3 represents the image's three-channel (R,G,B), and H and W represent the image's height and width, respectively. We employ the ResNet structure to construct the branch network. Initially, a convolution layer with a convolution kernel size of 7×7 extracts the image's structure features on a large perceptual range. Subsequently, four ResNet blocks with a convolution kernel size of 3×3 at different levels sequentially extract structural features, and an average pooling operation with a step size of 2 is performed on the extracted structural features. Subsequently, four more structure feature extraction and average pooling operations are performed on ResNet blocks with different layers. Finally, the feature is encoded into a 128-dimensional vector using the fully connected layer as the structural difference vector of the image.

To enhance network learning efficiency, we utilize the image line feature as a prior knowledge of the image structure. Initially, the line feature of the original input image is extracted to acquire the structure prior map. In the upsampling and decoding layer, after four convolution operations with a convolution kernel size of 3×3 and deconvolution processing with a step size of 2, the structure feature map of 1×H × W output is obtained. This strategy can guide the network to focus on the feature-rich region of the image by reducing the discrepancy between the output structure feature map and the structure prior map.

In the structure loss function $L_{structure}$, m is the structure feature map, m' is the prior of structure map. The range of m and m' are (0, 1). In the loss calculation, the cross entropy loss in the two

Biomedical Optics EXPRESS

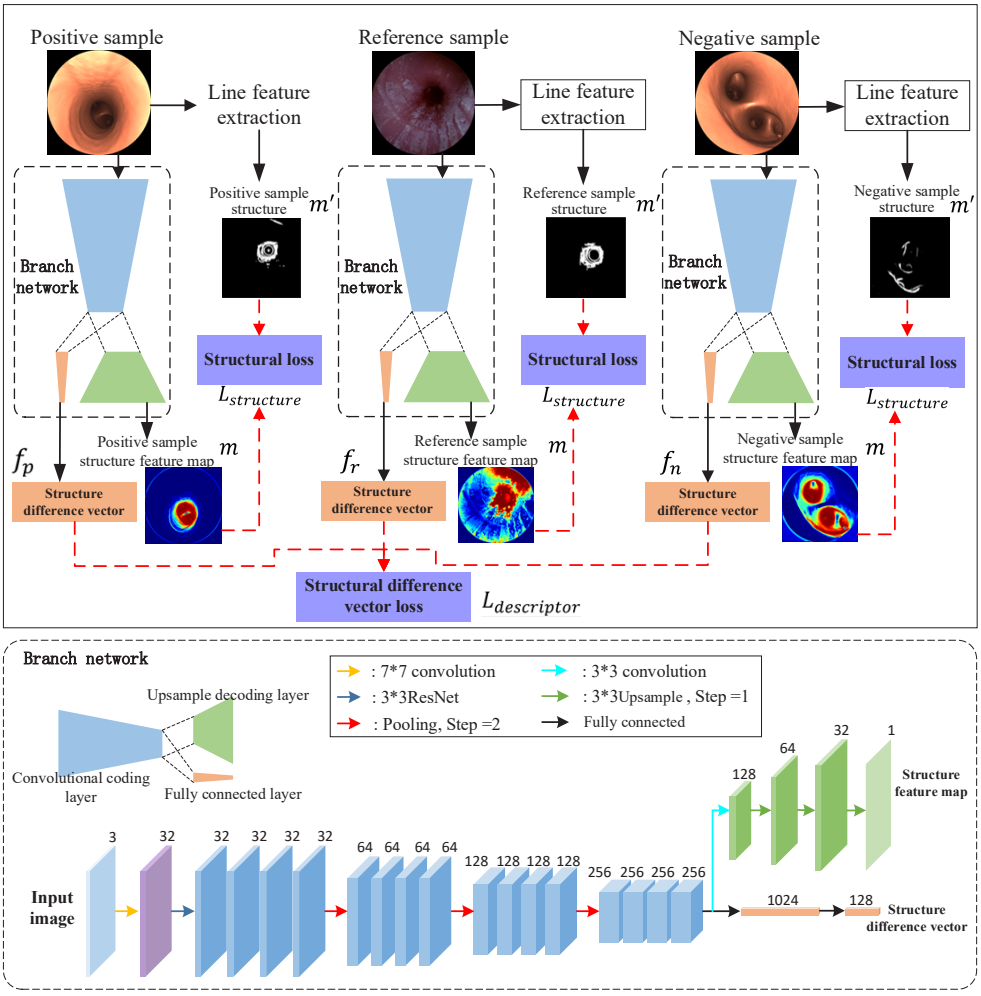

Fig. 2. The image structure difference (ISD) network architecture.

images is calculated pixel by pixel, and the loss matrix of the same size as the original image is obtained. The function is defined as follows:

$$L_{structure}(m, m') = -m\,logm' - (1-m)\,log(1-m') \tag{1}$$

To effectively cluster the structural difference vectors corresponding to each subregion image generated by the training network, we use the positive and negative sample comparison learning method to output the multi-branch clustering features from the model. Two outputs representing the regression offsets of positive samples and negative samples to reference images are generated through the fully connected layer. Therefore, the structural difference vector loss function L_{vector} is defined as follows:

$$L_{vector}(f_p, f_r, f_n) = max(\|f_p - f_r\| - \|f_n - f_r\| + \gamma, 0) \tag{2}$$

Where, f_p, f_r, f_n represents the image structure difference vectors of positive sample, reference sample and negative sample respectively. γ is the quantity of control boundary, and the adjustment range is $(0.5 \sim 2)$. By gradually reducing the distance between f_r and f_p, and increasing the

Biomedical Optics EXPRESS

distance between f_r and f_n, which is enables f_r obtain structure difference vector results with greater differentiation in different regional structures. Finally, the structure difference vector set $\{f_1, f_2, f_3, \ldots f_k\}$ of sequence frames was obtained.

The closest structural difference vector f_{min} is obtained by comparing the similarity between the structural difference vector of initial frame f_{new} and the sequence frame. This image location $p_{f_{min}\_location}$ corresponding to the structural difference vector is recorded as the initial frame position $p_{0\_location}$. The image structure difference vector similarity measure is defined as follows:

$$f_{min} \rightarrow \arg min \|f_{new} - f_i\|, f_i \in \{f_1, f_2, f_3, \ldots f_k\} \tag{3}$$

$$p_{0\_location} = p_{f_{min}\_location} \tag{4}$$

2.3. Pose estimation

In the virtual endoscopy image, the depth information of each frame can be obtained using the cavity model and the endoscopic pose. For the intraoperative endoscopy image, we utilized a self-supervised learning method to acquire image depth information [51]. Subsequently, we obtain the information entropy of the depth map of the virtual and intraoperative endoscopy images, respectively.

$$H' - \sum d'_{(i,j)} \log d'_{(i,j)} \tag{5}$$

$$\overline{H'}(s_n) = -\frac{1}{n} \sum_{(i,j)\in s_n}^{n} d'_{(i,j)} \log d'_{(i,j)} \tag{6}$$

$$H = -\sum d_{(i,j)} \log d_{(i,j)} \tag{7}$$

$$\bar{H}(s_n) = -\frac{1}{n} \sum_{(i,j)\in s_n}^{n} d_{(i,j)} \log d_{(i,j)} \tag{8}$$

Where, H' is the information entropy of the virtual endoscopy image depth map, H is the information entropy of the intraoperative endoscopy image depth map. s_n is the fan-shaped n equal division of the circular visual region of the endoscopy image. $\overline{H'}$ is the mean value of information entropy of the subblock region of the depth map of the virtual endoscopy image, $\bar{H}$ is the mean value of information entropy of the subblock region of the depth map of the intraoperative endoscopy image. Then, the similarity of the depth map information entropy between the intraoperative and virtual endoscopy images was measured by the cross-correlation mean square error similarity method of the image depth information entropy.

$$s_{corr\_info}(H, H') = \frac{1}{n} \sum \left((H - \bar{H}) - (H' - \overline{H'})\right)^2 \tag{9}$$

Where, s_{corr_info} is the similarity measurement method of cross-correlation mean square error. The motion variables Δp_{pose} of the two frames were estimated by maximizing the similarity measure of cross-correlation mean square error.

$$p_{k\_pose} = p_{(k-1)\_pose} + \Delta p_{pose} \tag{10}$$

$$\Delta p_{pose} = \arg \max_{\Delta p_{pose}} s_{corr\_info} \{H_k, H'(p_{(k-1)\_pose} + \Delta p_{pose})\} \tag{11}$$

Where, p_{k_pose} is the pose of p_k frame image, $p_{(k-1)\_pose}$ is the pose of $p_{(k-1)}$ frame image. This method estimated the optimal pose change Δp_{pose} by maximizing the depth map similarity between the intraoperative endoscopy image and the virtual endoscopic image of p_k frames.

Biomedical Optics EXPRESS

2.4. Repositioning optimization

The estimation of the interframe pose of the endoscopy image may produce drift errors, which gradually accumulate and seriously affect the positioning accuracy. To solve this problem, an adaptive boundary constraints method is used to optimize the position estimation results. To accurately calculate the error range of the current frame motion, we use the adaptive cavity boundary method to obtain the adaptive boundary constraint B in the cavity where the current frame p_c is located.

$$B(p_c) = F(p_c, \pi r^2) \tag{12}$$

$$r = \min \|g(p_c), s(p_c)\| \tag{13}$$

Where, r is the radius of the tangent circle. $g(p_c)$ is the point on the center line of the cross section of the cavity at the position of the current frame p_c. $s(p_c)$. is the cross section surface of the cavity at the position of the current frame p_c. By calculating the order of distance between point $g(p_c)$ and any point on the cross section surface $s(p_c)$, the minimum value of distance is obtained as the error threshold. $F(\Delta, \Delta)$ is the function to obtain the adaptive boundary constraint. Then, the adaptive error range was determined based on the adaptive boundary constraint conditions.

$$D_{relocation}(p_c) = \begin{cases} p_{c\_pose} \notin B(p_c), & 1 \\ p_{c\_pose} \in B(p_c) \ , & 0 \end{cases} \tag{14}$$

Were, $D_{relocation}$ is the determination method of the inclusion relationship between the current frame pose information and adaptive boundary constraints. If the pose of the current frame

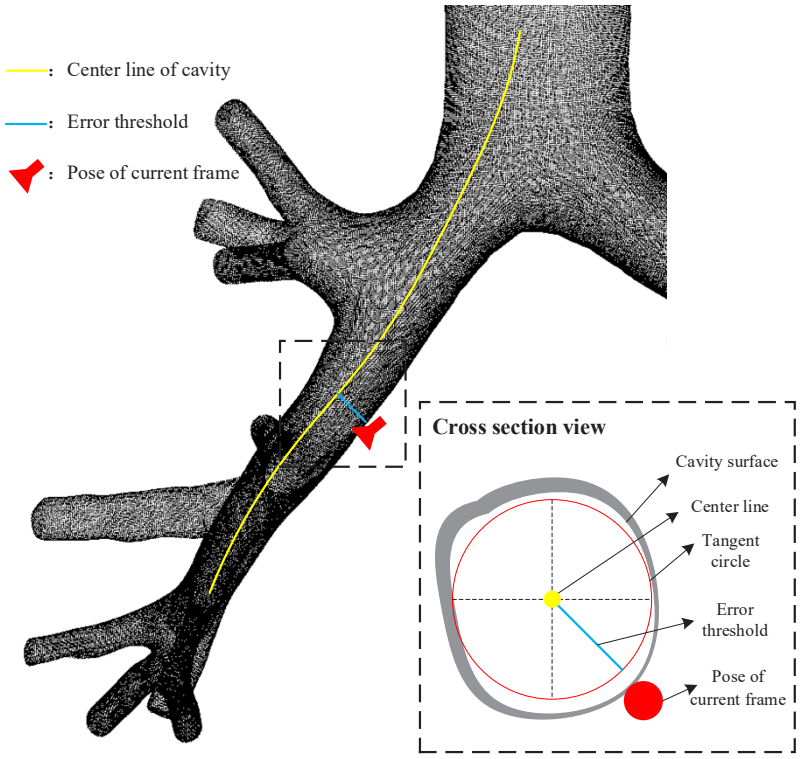

Fig. 3. The diagram of current frame position error determination.

Biomedical Optics EXPRESS

exceeds the adaptive boundary constraints, it indicates that the accumulated motion error of the current frame is too large, and the pose estimation should take the current frame as a new start frame. Otherwise, reposition optimization is not performed.

To prevent the cumulative error in the localization of multi-frame endoscopy in narrow cavities, an adaptive boundary constraints method is used for optimization, as illustrated in Fig. 3. In an independent branch structure of bronchial data, the yellow line represents the motion trajectory formed by the endoscopy location results of continuous frames, the red endoscopy denotes the current frame, and the red circle is inscribed on the structure. Initially, during the continuous frame endoscopy motion process, the cavity's spatial environment changes significantly, so we measure the minimum radius of the inscribed circle on the internal surface. Thereafter, the minimum radius is employed as the adaptive error threshold to determine relocation. Finally, if the error is greater than the threshold, the frame is utilized as the initial frame for relocation, and the optimization outcome of endoscope positioning is obtained.

3. Experiments

3.1. Datasets and implementation details

Our training data are generated from two unlabeled endoscopy videos, the first is from the Monocular Frames for Bronchoscopy Navigation (MBN) dataset [23], and the other is from the Hamlyn dataset [48]. In the MBN dataset, there are ten distinct scenarios datasets of surface, and the image resolution is 512*512. The ground truth of localization was recorded using virtual endoscopy roaming in the CT reconstruction model. In Hamlyn dataset, there are five different scenario datasets of surface, and the image resolution is 1280*1024. The ground truth of endoscopy localization was recorded with a robotic arm. The training set comprises 12 (8 from the MBN dataset and 4 from the Hamlyn dataset) data, while the validation set encompasses 3 (2 from the MBN dataset and 1 from the Hamlyn dataset) data, as demonstrated in Table 1.

Table 1. Dataset details of endoscopy image

Dataset	Category	Ground truth	Describe
MBN [23]	Bronchoscopy	Virtual endoscopy roaming in CT reconstruction model.	Contains 10 different scenario datasets of surface, image resolution is 512*512.
Hamlyn [48]	Colonoscopy	A robotic arm records localization.	Contains 5 different scenario datasets of internal surface, image resolution is 1280*1024.

To improve the robustness of the training network, several rendering methods are employed to supplement the image to simulate the localization ability with different lighting and texture changes. Figure 4 depicts the rendering images for different textures. The virtual endoscopy images were utilized to simulate the color and texture characteristics of the complex intraoperative environment to enhance the model's adaptability to different texture images. Figure 4 (1) Original image, (2) Darkened illumination processing, (3) Highlight removal processing, (4) Increased illumination processing, (5) Grayscale processing, (6) Noise addition processing, (7) Texture smoothing processing, (8) Depth map ground truth for the scene.

All experiments are conducted on a workstation with NVIDIA GeForce RTX 3090 GPU, with 24 GB memory. Pytorch framework is used to build the learning network. In depth information location estimation network, the learning rate of the optimization controls the updating rate of the weights, we use $d_{\alpha} = 0.001$ to get better convergence performance with faster initial learning efficiency. The exponential decay rate of the first moment estimate is $d_{beta1} = 0.9$. The

Biomedical Optics EXPRESS

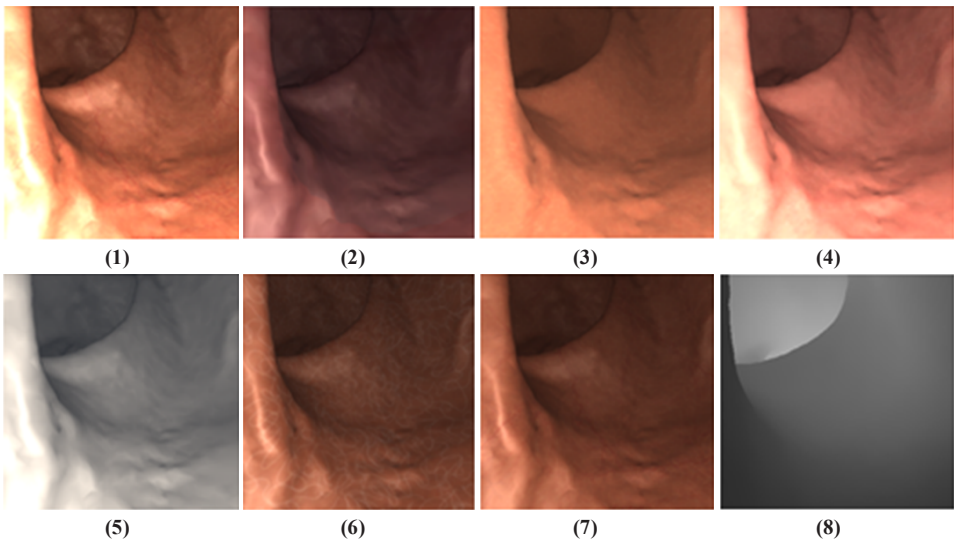

Fig. 4. Model extension training images. (1)-(7) show the rendered images for different textures, (8) depicts the depth map.

exponential decay rate of the second moment estimate is $d_{beta2} = 0.999$. The positive parameter factor is $d_\epsilon = 10E - 8$.

3.2. Evaluation metrics

The performance of the endoscopy localization method was evaluated using multiple indicators. The absolute translational error (ATE) can directly reflect the estimation accuracy and global consistency of trajectory. The relative posture error (RPE) describes the difference between the estimated position and the ground truth. Because the estimated pose and the ground truth pose are not in the same coordinate system, the iterative closest point (ICP) method was used to register the two poses. For the monocular endoscopy scale is uncertain, we use the similarity transformation matrix S to calculate the conversion from the estimated pose P to the ground truth Q, where the ATE and RPE of frame i in the estimated pose are defined as f_i and e_i respectively:

$$f_i = Q_i^{-1} S P_i \tag{15}$$

$$e_i = (Q_i^{-1} Q_{i+1})^{-1} (P_i^{-1} P_{i+1}) \tag{16}$$

Then, the root mean square error (R.SE) was used to make statistics on ATE and RPE of each frame, and an overall evaluation index was obtained respectively:

$$F_{RMSE}(f_{i:n}, \Delta) \left(\frac{1}{m} \sum_{i=1}^{m} \|f_i\|^2 \right)^{\frac{1}{2}} \tag{17}$$

$$E_{RMSE}(e_{i:n}, \Delta) = \left(\frac{1}{m} \sum_{i=1}^{m} \|e_i\|^2 \right)^{\frac{1}{2}} \tag{18}$$

In initial frame localization, the proportion of correct estimated is used to evaluate the performance of the algorithm. To evaluate the accuracy of the results, using the following

Biomedical Optics EXPRESS

formula:

$$Accuracy = \frac{Q_t}{Q_N} \times 100\% \tag{19}$$

Where, Q_t. is the number of correctly segmented estimated image, and Q_N is the total number of images.

3.3. Initial frame location estimation

Virtual endoscopy roaming was utilized to simulate the movement process of actual endoscopy in model data. The topology of the bronchus was rich and the path was narrow, and the rotation angle at the connection was large, which could lead to significant errors in localization and affect the segmented estimation. Compared with the model data, real endoscopy data can more effectively reflect the uncertainties of clinical operation. For example, doctors may need to repeat their observation of the same region during the operation, leading to back-and-forth movement of the endoscope. Figure 5 illustrates two groups of bronchoscopy data and two groups of colonoscopy data used to compare the accuracy of segmented localization estimation between the Offset-Net [49] method and the proposed method. Figures (1) and (2) denote the results of the Offset-Net method and the proposed method, respectively. Figures (a) and (b) are two groups of bronchoscopy results, while (c) and (d) are two groups of colonoscopy results. The blue dots represent the correct segment frames, whereas the red dots represent the incorrect segment frames in segment estimation. The yellow lines denote the ground truth localization of the endoscopy. In bronchoscopy data, the Offset-Net method has a significant cumulative impact on the estimation errors, specifically in the region with a larger rotation perspective. In colonoscopy data, the Offset-Net method tends to confuse the estimation position in the turn-back scenario, resulting in several initial frame estimation errors. Our method can maintain a high level of correct estimation in various data scenarios and has a high degree of robustness.

The number of error frames in continuous frames reflects the reliability of initial positioning results. We employed the ratio of the quantity of errors (Our / Offset-Net) as the evaluation index, as indicated in Table 2. In the experiment, six groups of bronchoscopy data and four groups of colonoscopy data were utilized for segmented location estimation, and each scene was divided into the front, middle, and end segments for comparison based on the average number of total frames. As Table 2 indicates, the number of estimation errors gradually increases in the Offset-Net method as the endoscopy moves, making it more susceptible to adverse factors during endoscopy shooting. Our method can maintain a low number of estimation errors in various data scenes and has high robustness towards various influencing factors.

Table 2. Average estimation results of initial frame location

Dataset	Scene	Front	Middle	End
Bronchoscopy	1	0.41	0.23	0.56
	2	0.63	0.33	0.18
	3	0.56	0.33	0.25
	4	0.59	0.50	0.29
	5	0.37	0.56	0.26
	6	0.42	0.38	0.49
Colonoscopy	1	0.60	0.50	0.38
	2	0.27	0.31	0.08
	3	0.55	0.29	0.17
	4	0.43	0.30	0.34

Biomedical Optics EXPRESS

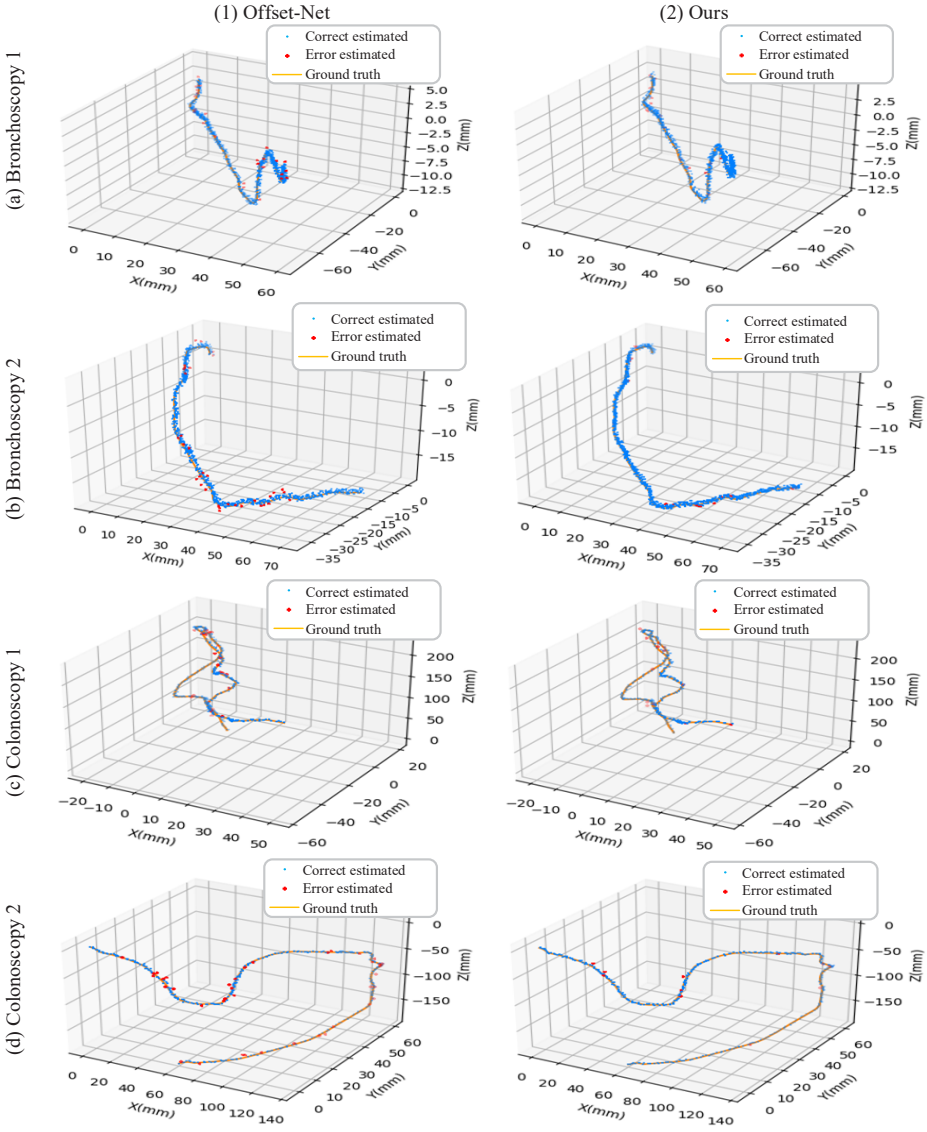

Fig. 5. Initial frame location estimation results. (1) and (2) denote the results of the Offset-Net and the proposed method, (a) and (b) are two groups of bronchoscopy results, (c) and (d) are two groups of colonoscopy results.

Biomedical Optics EXPRESS

Table 3 compares the accuracy of the Offset-Net and our method, where the ISD network was employed to estimate the initial frame location. In bronchoscopy and colonoscopy data, the average accuracy reached 95.8% and 95.7%, respectively, which is 2.7% and 5.4% higher than that of the Offset-Net method. The experimental outcomes indicate that our method outperforms the Offset-Net method in six scenarios in terms of segmented location estimation results. In bronchoscopy data, scene 3 contains motion blur, and topological branching leads to a decrease in the accuracy of the estimation results. In colonoscopy data, our method can still obtain high accuracy estimation outcomes.

Table 3. Accuracy comparison of initial frame location estimation

Method	Bronchoscopy				Colonoscopy	
Scene	1	2	3	4	1	2
Offset-Net	89.2%	94.2%	91.9%	97.2%	91.4%	89.3%
Ours	94.8%	96.5%	93.9%	98.1%	96.4%	95.0%

3.4. *Interframe pose estimation*

We verified the angle error and translation error of positioning estimation of 1000 continuous frame endoscope images by comparing them with three different position methods, as depicted in Fig. 6. (a) shows the error of the angle result, and (b) represents the error of the translation result. The yellow line indicates the Offset-Net method's outcome, the green line represents the Iopt-Net [50] method, the blue line corresponds to the HMT [31] method, and the red line shows the result of our method. Concerning the error of angle results, the Iopt-Net and HMT methods have a larger range of error and lower estimation stability in continuous frame location estimation results. The Offset-Net method has high accuracy in the initial estimation, but the error increases as the frames advance continuously. The error range rises abruptly due to the cumulative error after 400 frames. Our method has a high initial error, but the error returns to a lower range and remains stable with the process of continuous frame advance. Regarding the error of the translation result, the HMT method has a larger range of errors. The Iopt-Net method can keep the error range low and stable. The Offset-Net method has high accuracy in the initial estimation, but the error increases with the process of continuous frame advance. The error range increases sharply due to the cumulative error after 200 frames. Our method can remain stable with the process of continuous frame advance. Overall, our method exhibits high estimation stability in the two error indicators and has high robustness in continuous frame location estimation.

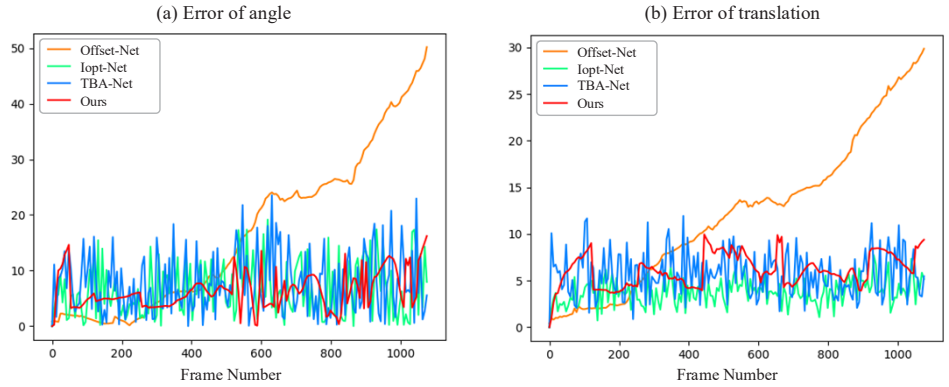

Fig. 6. Interframe pose estimation performance. (a) represents the error of the angle result, (b) represents the error of the translation result.

Biomedical Optics EXPRESS

We also quantitatively evaluated the location estimation of continuous frame using six groups of bronchoscopy data and four groups of colonoscopy data. We compared the results of the HMT, Offset-Net, and Iopt-Net methods, as indicated in Table 4. The absolute trajectory error (ATE) and relative attitude error (RPE) were used to estimate the location and ground truth. The HMT method combines the epipolar constraints with the Kalman filter and image registration technique to obtain the estimation of the location displacement between frames. However, the accuracy of this method significantly decreases in ATE and RPE due to the influence of image quality and tissue deformation. The Offset-Net method uses the generative adversarial network to locate the endoscopy in different environments, but in the data with a long motion period, the error of the two indicators is large. On the other hand, the Iopt-Net method can effectively avoid the cumulative error of multi-frame location by combining the two-path ResNet structure with the iterative optimization algorithm. In the bronchoscopy data, the second group of RPE and the fifth group of ATE indicators obtained better results. Our method reduced the errors by 17.9%, 28.5%, and 15.2%, 19.5% in the two indicators. Overall, the comprehensive experiments demonstrated that our method has good performance in location estimation in various scenarios.

Table 4. Comparison of ATE and RPE with continuous frames.

Dataset	Scene	Evaluation	HMT	Offset-Net	Iopt-Net	Ours
Bronchoscopy	1	ATE	2.03	1.23	0.59	0.38
		RPE	0.87	0.29	0.17	0.16
	2	ATE	6.52	3.26	2.37	2.04
		RPE	0.77	0.48	0.51	0.59
	3	ATE	9.13	7.04	2.62	1.56
		RPE	0.79	0.44	0.35	0.26
	4	ATE	3.74	1.96	0.88	0.57
		RPE	0.71	0.28	0.45	0.25
	5	ATE	5.16	3.83	2.14	2.57
		RPE	0.75	0.24	0.54	0.21
	6	ATE	4.45	2.84	1.51	1.47
		RPE	0.72	0.33	0.64	0.32
Colonoscopy	1	ATE	4.27	2.75	1.48	1.21
		RPE	0.79	0.56	0.38	0.29
	2	ATE	7.33	5.15	2.49	1.76
		RPE	0.61	0.55	0.43	0.38
	3	ATE	4.45	2.73	1.52	1.48
		RPE	0.68	0.52	0.47	0.36
	4	ATE	4.51	3.74	1.82	1.62
		RPE	0.74	0.53	0.61	0.49

3.5. Positioining the optimization result

Continuous frame endoscopy location estimation is susceptible to cumulative errors. To address this issue, we adopted the adaptive boundary constraints method to acquire position updates, as depicted in Fig. 7. We compared the localization estimation accuracy between the Offset-Net method and the proposed method using two groups of bronchoscopy data and two groups of colonoscopy data. (1) shows the results of the Offset-Net method, while (2) shows the proposed method. (a) and (b) display two groups of bronchoscopy results, whereas (c) and (d) represent two groups of colonoscopy results. The blue line represents the optimization outcome of continuous

Biomedical Optics EXPRESS

frame location estimation or adaptive boundary constraints estimation, while the yellow line corresponds to the ground truth. In the experiment's four groups of data outcomes, the initial estimation results of continuous frame localization are accurate. However, the cumulative error significantly increases with multi-frame location estimation, particularly in the rotating region and turn-back path region. The adaptive boundary constraints method can efficiently suppress cumulative error and enhance positioning accuracy.

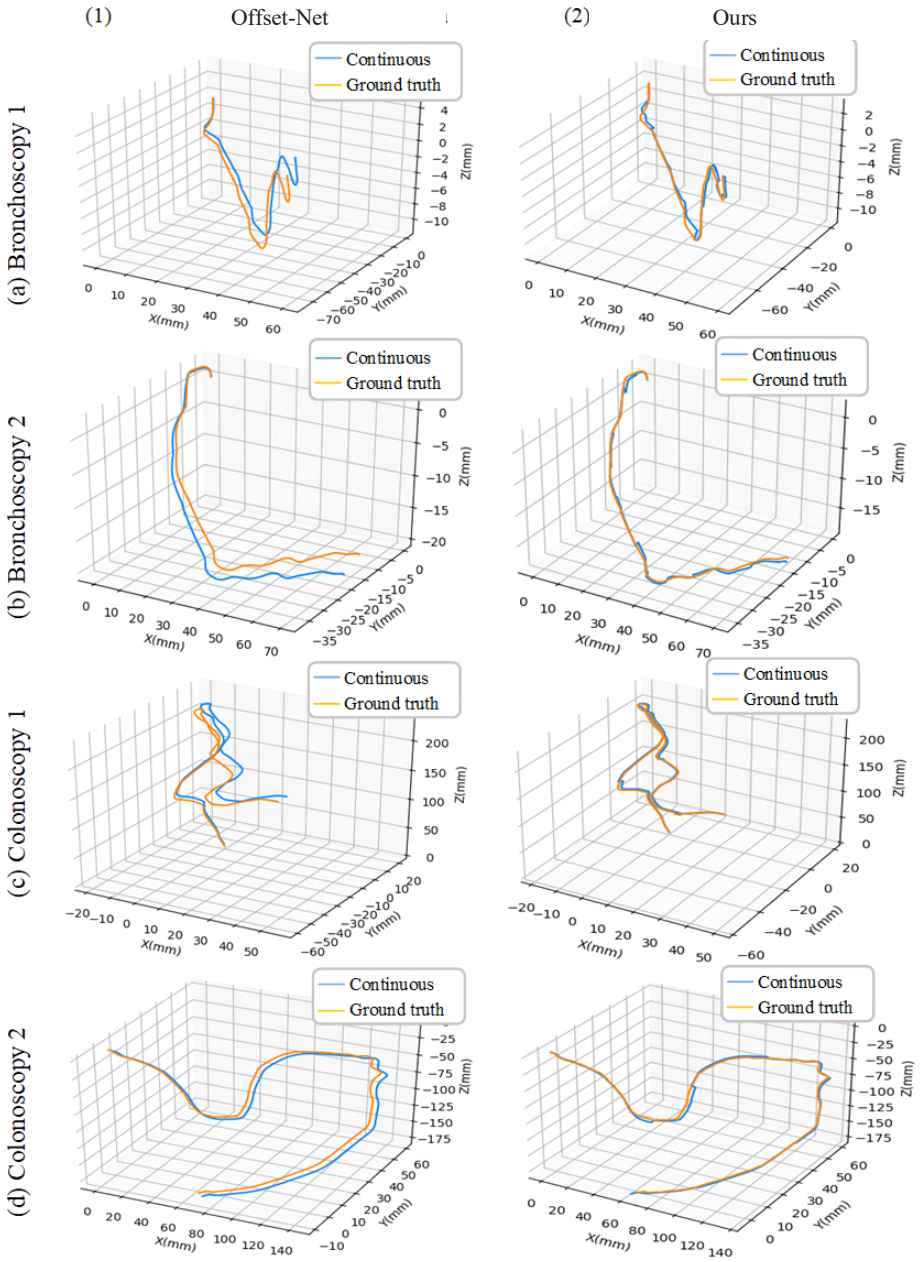

Fig. 7. Repositioning optimization results. (1) and (2) shows the results of the Offset-Net and ours method. (a) and (b) display two groups of bronchoscopy results, (c) and (d) represent two groups of colonoscopy results.

Biomedical Optics EXPRESS

Table 5. Time efficiency comparison of different methods.

Dataset	Method	Error (mm)			Time (s)		
		mean	SD	max	mean	SD	max
Bronchoscopy	Offset-Net	1.68	1.69	1.99	0.82	2.49	1.05
	HMT	3.05	2.52	3.29	12.01	2.92	20.84
	Iopt-Net	1.59	1.26	0.99	1.76	3.63	6.37
	Ours	1.43	1.61	1.82	0.93	1.01	1.19
Colonoscopy	Offset-Net	5.26	3.62	6.75	1.23	1.16	2.45
	HMT	14.59	5.18	19.95	15.92	3.95	29.92
	Iopt-Net	5.04	1.81	5.63	2.34	2.83	8.97
	Ours	3.64	1.46	3.77	1.35	1.36	3.55

In Fig. 8, we present the continuous frame endoscopy localization estimation results of different topologies on the same bronchial data. (a)-(c) demonstrate the localization results of the three branches. The black wireframe displays an enlarged view of branch details, the yellow line represents the ground truth, and the red dot denotes the endoscope positioning estimation results in each frame. Experimental outcomes indicated that our proposed method effectively avoids the confusion of similarity in different topological structure estimations of the same data and maintains high robustness in varying scenarios.

To evaluate our method's performance, we compared its location error and time consumption with three other methods. Six groups of bronchoscopy data and four groups of colonoscopy data were sampled and verified in the experiment. To enable a fair comparison, we normalized the location error and time consumption and compared them in the same order of magnitude. The results are presented in Table 5. Among the bronchoscopy and colonoscopy data, Offset-Net demonstrated excellent time efficiency. The HMT method had high location error and low time efficiency. The Iopt-Net method offered excellent stability. Our method achieved the lowest location error and similar stability to the Iopt-Net. Moreover, our approach met the real-time performance requirements in various scenarios.

By statistically analyzing the mean and variance of the number of error frames, this chapter's method can reflect the precision changes in the localization estimation. In the experiment, 6 groups of bronchial images and 4 groups of intestinal images were used, and they were divided into the front, middle, and back segments according to the average total number of frames for comparison, as shown in Fig. 9. The blue line in the figure represents the mean number of localization error frames for bronchial images, the orange line represents the mean number of localization error frames for intestinal images, the gray line represents the variance of the number of localization error frames for bronchial images, and the yellow line represents the variance of the number of localization error frames for intestinal images. The changing trends show that the mean number of localization error frames for both bronchial and intestinal images is decreasing, indicating that the Offset-Net method has a continuously decreasing localization accuracy compared to the method in this chapter. The variance of the number of localization error frames for both bronchial and intestinal images shows a stable trend, indicating that the method in this chapter is not easily affected by changes in the number of images and has good localization accuracy stability.

Biomedical Optics EXPRESS

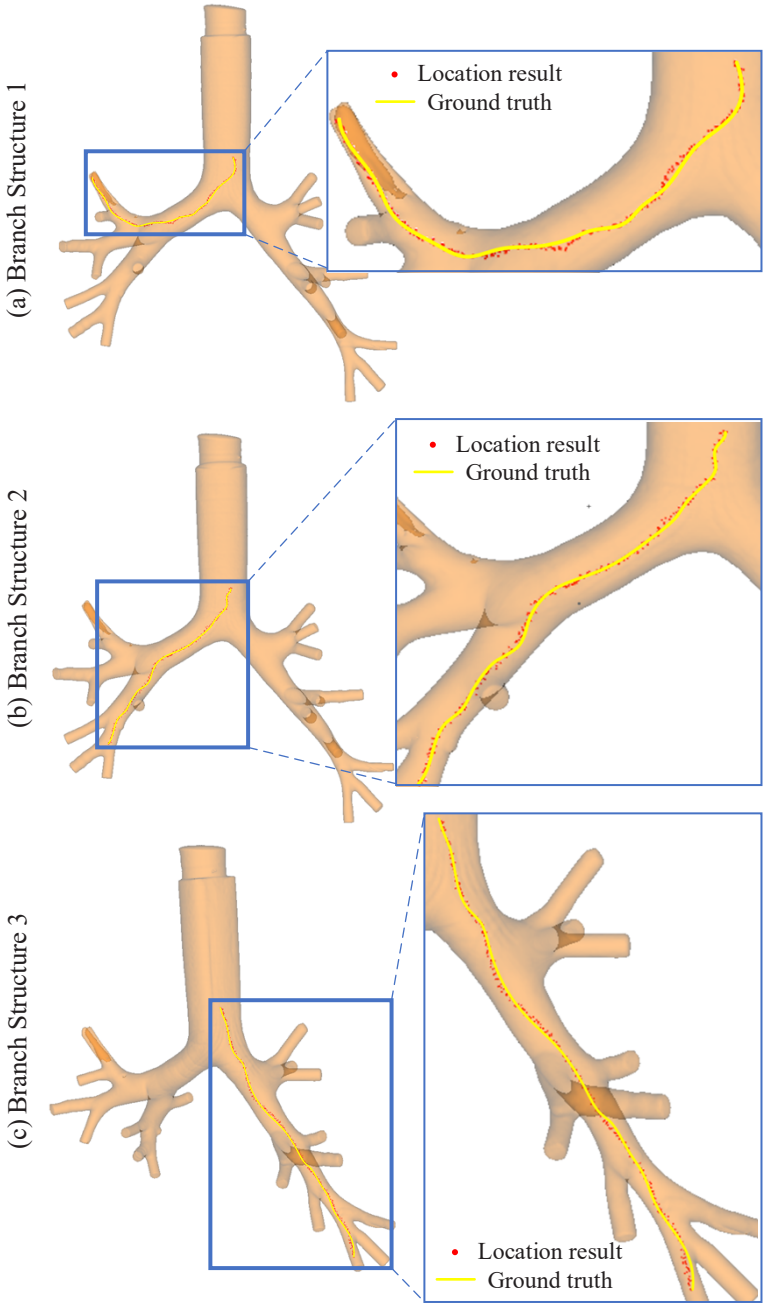

Fig. 8. Results of endoscopy image pose estimation. (a)-(c) demonstrate the localization results of the three branches.

Biomedical Optics EXPRESS

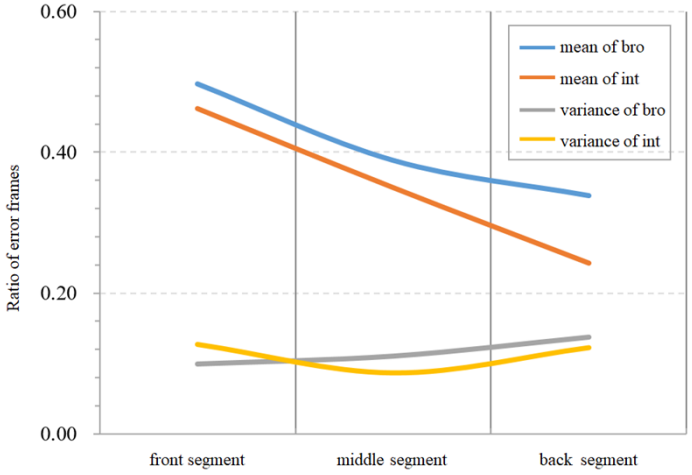

Fig. 9. Error analysis of pose estimation in different scenarios.

4. Conclusion and discussion

This work presents a structure-depth information based monocular endoscopy localization method, which is used to improve the position ability of endoscopy based on the 2D images. The ISD network was used to generate the structure difference vector of each endoscopic image. To improve the efficiency of network learning, the prior knowledge of image structure is used to guide the network focus region in an image. An image depth information entropy similarity maximization method is introduced to improve the accuracy of pose estimation. Then, the repositioning strategies are used to avoid cumulative errors affecting the results. Experiments with published datasets show that the proposed method can improve the accuracy of pose estimation of monocular endoscopy images and reconstruct more accurate endoscopy motion at the same time.

Funding. National Key Research and Development Program of China (2023YFC2415300); National Natural Science Foundation of China (62025104, 62171039, 62202045); Beijing Municipal Natural Science Foundation (L222149).

Disclosures. The authors declare that there are no conflicts of interest related to this article.

Data Availability. Data availability. Data underlying the results presented in this paper are available in Ref. [23] and [48].

References

1. N. Mahmoud, T. Collins, A. Hostettler, *et al.*, "Live tracking and dense reconstruction for handheld monocular endoscopy," IEEE Trans. Med. Imaging **38**(1), 79–89 (2019).
2. D. Than T, G. Alici, H. Zhou, *et al.*, "A review of localization systems for robotic endoscopic capsules[J]," IEEE Trans. Biomed. Eng. **59**(9), 2387–2399 (2012).
3. J. Oh Y, G. Yang S, H. Han W, *et al.*, "Effectiveness of Intraoperative Endoscopy for Localization of Early Gastric Cancer during Laparoscopic Distal Gastrectomy[J]," Dig. Surg. **39**(2-3), 92–98 (2022).
4. J. Kim, H. Al Faruque, S. Kim, *et al.*, "Multimodal endoscopic system based on multispectral and photometric stereo imaging and analysis[J]," Biomed. Opt. Express **10**(5), 2289–2302 (2019).
5. J. Bian, Z. Li, N. Wang, *et al.*, "Unsupervised scale-consistent depth and ego-motion learning from monocular video[J]," Adv. neural inform. Process. Syst. **32** (2019).
6. C. Xie, T. Yao, J. Wang, *et al.*, "Endoscope localization and gastrointestinal feature map construction based on monocular slam technology[J]," Journal of infection and public health **13**(9), 1314–1321 (2020).
7. R Widya A, Y Monno, M Okutomi, *et al.*, "Learning-based depth and pose estimation for monocular endoscope with loss generalization[C]," *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 2021: 3547–3552.
8. C. Wang, Y. Hayashi, M. Oda, *et al.*, "Depth-based branching level estimation for bronchoscopic navigation[J]," Int J CARS **16**(10), 1795–1804 (2021).
9. E. Spyrou and K. Iakovidis D, "Video-based measurements for wireless capsule endoscope tracking[J]," Meas. Sci. Technol. **25**(1), 015002 (2014).

Biomedical Optics EXPRESS

10. S Shao, Z Pei, W Chen, *et al.*, "Self-supervised learning for monocular depth estimation on minimally invasive surgery scenes[C]," *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021: 7159–7165.
11. S Rajendra, J Nine, S Saleh, *et al.*, "Towards End-to-End Estimation of Camera Trajectory With Deep Monocular Visual Odometry[C]," *International Symposium on Computer Science, Computer Engineering and Educational Technology (ISCSET-2021)*. 88.
12. L Qiu and H. Ren, "Endoscope navigation and 3D reconstruction of oral cavity by visual SLAM with mitigated data scarcity[C]," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018: 2197–2204.
13. K. B. Ozyoruk, G. I. Gokceler, T. L. Bobrow, *et al.*, "EndoSLAM dataset and an unsupervised monocular visual odometry and depth estimation approach for endoscopic videos[J]," Med. Image Anal. **71**, 102058 (2021).
14. G. Dimas, E. Spyrou, K. Iakovidis D, *et al.*, "Intelligent visual localization of wireless capsule endoscopes enhanced by color information[J]," Comput. Biol. Med. **89**, 429–440 (2017).
15. S. Leonard, A. Sinha, A. Reiter, *et al.*, "Evaluation and stability analysis of video-based navigation system for functional endoscopic sinus surgery on in vivo clinical data[J]," IEEE Trans. Med. Imaging **37**(10), 2185–2195 (2018).
16. T. Feng and D. Gu, "SGANVO: Unsupervised deep visual odometry and depth estimation with stacked generative adversarial networks[J]," IEEE Robot. Autom. Lett. **4**(4), 4431–4437 (2019).
17. J. Herp, U. Deding, M. Buijs M, *et al.*, "Feature point tracking-based localization of colon capsule endoscope[J]," Diagnostics **11**(2), 193 (2021).
18. S. Bernhardt, A. Nicolau S, V. Agnus, *et al.*, "Automatic localization of endoscope in intraoperative CT image: A simple approach to augmented reality guidance in laparoscopic surgery[J]," Med. Image Anal. **30**, 130–143 (2016).
19. S Mathew, S Nadeem, S Kumari, *et al.*, "Augmenting colonoscopy using extended and directional cyclegan for lossy image translation[C]," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020: 4696–4705.
20. Y. Ming, X. Meng, C. Fan, *et al.*, "Deep learning for monocular depth estimation: A review[J]," Neurocomputing **438**, 14–33 (2021).
21. L. Lurie K, R. Angst, V. Zlatev D, *et al.*, "3D reconstruction of cystoscopy videos for comprehensive bladder records[J]," Biomed. Opt. Express **8**(4), 2106–2123 (2017).
22. S. Lee, S. Shim, G. Ha H, *et al.*, "Simultaneous optimization of patient–image registration and hand–eye calibration for accurate augmented reality in surgery[J]," IEEE Trans. Biomed. Eng. **67**(9), 2669–2682 (2020).
23. M. Visentini-Scarzanella, T. Sugiura, T. Kaneko, *et al.*, "Deep monocular 3D reconstruction for assisted navigation in bronchoscopy[J]," Int J CARS **12**(7), 1089–1099 (2017).
24. K He, X Zhang, S Ren, *et al.*, "Deep residual learning for image recognition[C]," *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: 770–778.
25. V. Penza, A. S. Ciullo, S. Moccia, *et al.*, "Endoabs dataset: Endoscopic abdominal stereo image dataset for benchmarking 3d stereo reconstruction algorithms[J]," The International Journal of Medical Robotics and Computer Assisted Surgery **14**(5), e1926 (2018).
26. Y. Almalioglu, R. U. Saputra M, P. B. De Gusmao P, *et al.*, "GANVO: Unsupervised deep monocular visual odometry and depth estimation with generative adversarial networks[C]," *2019 International conference on robotics and automation (ICRA)*, IEEE, 5474–5480 (2019).
27. A. Rau, J. E. Edwards P, F. Ahmad O, *et al.*, "Implicit domain adaptation with conditional generative adversarial networks for depth prediction in endoscopy[J]," Int J CARS **14**(7), 1167–1176 (2019).
28. Y. Almalioglu, M. Turan, R. U. Saputra M, *et al.*, "SelfVIO: Self-supervised deep monocular Visual–Inertial Odometry and depth estimation[J]," Neural Networks **150**, 119–136 (2022).
29. A. Banach, F. King, F. Masaki, *et al.*, "Visually navigated bronchoscopy using three cycle-consistent generative adversarial network for depth estimation[J]," Med. Image Anal. **73**, 102164 (2021).
30. N Mahmoud, A Nicolau S, A Keshk, *et al.*, "Fast 3d structure from motion with missing points from registration of partial reconstructions[C]," *Articulated Motion and Deformable Objects: 7th International Conference*, AMDO 2012, Port d'Andratx, Mallorca, Spain, July 11-13, 2012. Proceedings 7. Springer Berlin Heidelberg, 2012: 173–183.
31. C. Zhao, L. Sun, Z. Yan, *et al.*, "Learning Kalman Network: A deep monocular visual odometry for on-road driving[J]," Robotics Autonomous Syst. **121**, 103234 (2019).
32. R. Mur-Artal, M. M. Montiel J, and D. Tardos J, "ORB-SLAM: a versatile and accurate monocular SLAM system[J]," IEEE Trans. Robot. **31**(5), 1147–1163 (2015).
33. A. Merritt S, R. Khare, R. Bascom, *et al.*, "Interactive CT-video registration for the continuous guidance of bronchoscopy[J]," IEEE Trans. Med. Imaging **32**(8), 1376–1396 (2013).
34. R. Ma, R. Wang, Y. Zhang, *et al.*, "RNNSLAM: Reconstructing the 3D colon to visualize missing regions during a colonoscopy[J]," Med. Image Anal. **72**, 102100 (2021).
35. L. Chen, W. Tang, W. John N, *et al.*, "SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality[J]," Comput. Methods Programs Biomed. **158**, 135–146 (2018).
36. F. Bardozzo, T. Collins, A. Forgione, *et al.*, "StaSiS-Net: A stacked and siamese disparity estimation network for depth reconstruction in modern 3D laparoscopy[J]," Med. Image Anal. **77**, 102380 (2022).
37. K. İncetan, I. O. Celik, A. Obeid, *et al.*, "VR-Caps: a virtual environment for capsule endoscopy[J]," Med. Image Anal. **70**, 101990 (2021).

Biomedical Optics EXPRESS

38. S. Zhang, L. Zhao, S. Huang, *et al.*, "A template-based 3D reconstruction of colon structures and textures from stereo colonoscopic images[J]," IEEE Trans. Med. Robot. Bionics **3**(1), 85–95 (2021).
39. H Lim, Y Kim, K Jung, *et al.*, "Avoiding degeneracy for monocular visual SLAM with point and line features[C]," *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021: 11675–11681.
40. M. Shen, Y. Gu, N. Liu, *et al.*, "Context-aware depth and pose estimation for bronchoscopic navigation[J]," IEEE Robot. Autom. Lett. **4**(2), 732–739 (2019).
41. F. Mahmood and J. Durr N, "Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy[J]," Med. Image Anal. **48**, 230–243 (2018).
42. D. Recasens, J. Lamarca, M. Fácil J, *et al.*, "Endo-depth-and-motion: Reconstruction and tracking in endoscopic videos using depth networks and photometric constraints[J]," IEEE Robot. Autom. Lett. **6**(4), 7225–7232 (2021).
43. X. Ban, H. Wang, T. Chen, *et al.*, "Monocular visual odometry based on depth and optical flow using deep learning[J]," IEEE Trans. Instrum. Meas. **70**, 1–19 (2021).
44. S. Shao, Z. Pei, W. Chen, *et al.*, "Self-supervised monocular depth and ego-motion estimation in endoscopy: Appearance flow to the rescue[J]," Med. Image Anal. **77**, 102338 (2022).
45. L. Lei, J. Li, M. Liu, *et al.*, "Shape from shading and optical flow used for 3-dimensional reconstruction of endoscope image[J]," Acta Oto-Laryngol. **136**(11), 1190–1192 (2016).
46. E Hoffer and N. Ailon, "Deep metric learning using triplet network[C]," *Similarity-Based Pattern Recognition: Third International Workshop, SIMBAD 2015*, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3. Springer International Publishing, 2015: 84–92.
47. G. Huang, Z. Liu, L. Van Der Maaten, *et al.*, "Densely connected convolutional networks[C]," *Proceedings of the IEEE conference on computer vision and pattern recognition.*, 4700–4708 (2017).
48. Imperial College London, "Hamlyn Centre Laparoscopic/Endoscopic Video Datasets," Imperial College London, 2023, http://hamlyn.doc.ic.ac.uk/vision/.
49. J. Sganga, D. Eng, C. Graetzel, *et al.*, "Offsetnet: Deep learning for localization in the lung using rendered images," *2019 international conference on robotics and automation (ICRA)*, IEEE, 5046–5052 (2019).
50. J. Song, M. Patel, A. Girgensohn, *et al.*, "Combining deep learning with geometric features for image-based localization in the Gastrointestinal tract[J]," Expert Syst. Appl. **185**, 115631 (2021).
51. X. Liu, A. Sinha, M. Ishii, *et al.*, "Dense depth estimation in monocular endoscopy with self-supervised learning methods[J]," IEEE Trans. Med. Imaging **39**(5), 1438–1447 (2020).